Final Report (Second Year Grant)

Grant AOARD-11-4056

"Effects of Cognitive Load on Trust"
May 2012

Fang Chen
NICTA

<u>Executive Summary</u>

This report summarizes the research activities undertaken as part of the "Effects of Cognitive Load on Trust" project in conjunction with the US AFRL and Sunway University, Malaysia. NICTA's role comprised the measurement and assessment of cognitive load through speech and other interaction modalities. The project is focused on the examination of the relationship between cognitive load and trust judgments, and the effect of cultural differences in the way trust judgments are made.

The second year of the project has been dedicated to the analysis of the Australian dataset, collected in 2011, and the second data collection phase from the US and Malaysian sites. A multidimensional data analysis was planned to analyze various modality data collected including subjective ratings, speech signal data, linguistic data, and interaction data (both mouse and keyboard interactions) and their behavior under different cognitive load conditions. The primary outcomes for this part of the work are described in the first part of this report, and a summary of the data collection outcomes so far is included in the second part of the report.

Report Documentation Page

1. REPORT DATE **25 MAY 2012**	2. REPORT TYPE **Final**	3. DATES COVERED **09-05-2012 to 08-05-2012**

4. TITLE AND SUBTITLE **Effects of Cognitive Load on Trust**	5a. CONTRACT NUMBER **FA23861114056**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Fang Chen**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **NICTA,13 Garden Street,Eveleigh NSW 2015,Australia,AU,Australia**	8. PERFORMING ORGANIZATION REPORT NUMBER **N/A**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **AOARD, UNIT 45002, APO, AP, 96338-5002**	10. SPONSOR/MONITOR'S ACRONYM(S) **AOARD**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AOARD-114056**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This is the final report of a project that was focused on the examination of the relationship between cognitive load and trust judgments, and the effect of cultural differences in the way trust judgments are made.

15. SUBJECT TERMS
Cognitive Psychology , Behavioural Science

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

Table of Contents

1. Introduction

Trust is found to be a critical factor driving human behavior in both interpersonal and computer-based interactions. Previous research by Mayer et al. [1] has found three trustworthiness elements that influence the development of trust in interpersonal situations: ability, benevolence, and integrity. Thus far, only a few studies have looked at how different situational factors influence trust development as reflected in the relative salience of the three trustworthiness indicators. One dominant situational factor that may shape trust perceptions of an information source is culture. Similarly, little is known how cognitive load may affect the different trustworthiness factors during trust development and acquisition.

The 3-year research project proposed serves as part of a larger international research effort in collaboration with Dr. Lyons and Dr. Stokes (AFRL), and Dr. Yeo (University of Malaysia Sarawak), with separate proposals to be submitted through the AFOSR/AOARD programs. A three-part user experiment was designed - one in the US, one in Australia, and one in Malaysia, to investigate the cross-cultural influences on trust. The Australian part of the data collection was completed in 2011 and its analysis is in progress thus far, with the US and Malaysian data collection currently in progress.

2. Project Plan Updates

The following project plan was agreed to as part of the grant approval for the second year of this project. In 2011, milestones 1 – 4 (shaded in grey) were completed. This year, milestones 5-10 have been amended from previous documents, as some analysis of interactive features had not been included in the original proposal. Additionally, a team member from NICTA left the project and was replaced; hence the project's timeline was affected. As of the time of writing, the project is running on time as per schedule.

ID	Milestone	Deliverable/Outcome	Due Date
M1	Complete Pre-Pilots (Materials)	• Pilot test the neutrality of the stimulus data to be used in the experiment • Stimulus material in target demographic (Australian) • Make changes to the stimulus material as appropriate to ensure neutrality	Jul 30, 2011
M2	Experiment Tool Design	• Development of the experimental application to be used • Implement factor manipulations, including cognitive load • Implement data collection functionality as part of the design	Aug 31, 2011
M3	Complete Pre-Pilots (Study Design)	• Conduct pilots on target demographic (6 participants) • Evaluate study design, procedure,	Sep 30, 2011

		physical set-up • Assess changes needed at each site	
M4	Complete Experimental Study	• Source participants • Run the study • Debrief participants	Nov 30, 2011
M5	Linguistic Analysis of Speech Data	• Prepare speech transcriptions and annotations • Run linguistic analyses on text data derived from speech • Report results	July 31, 2012
M6	Signal Analysis of Speech Data	• Collect speech data from other sites • Segment, annotate and label speech data • Build speech models to represent cognitive load levels. • Report results to rest of the team	Sep 30, 2012
M7	Consolidate Speech/Linguistic Findings	• Ground truth analysis (subjective ratings, performance) • Contextualise the findings with those from Trust based manipulations, looking for interaction effects	Nov 30, 2012
M8	Interactive Data (Mouse movements) Analysis	• Develop features that may be affected by load • Build a parsing tool to extract relevant features • Statistical analysis of results (by load and trust components)	Feb 28, 2013
M9	Final Year Report	• Produce Final Year report on findings, data summaries and conclusions	May 31, 2013
M10	Project Management	• Weekly meetings • Team workshops, including conference calls with co-investigators • Year-end final report circulated to AOARD office and all other investigators	May 31, 2013

3. User Study Design and Materials

Hypotheses

A detailed literature review was conducted in 2011 to understand the state of the art in the trust and cognitive load domains (see our first year report for the review of the literature). Based on our review and as a first step to gain insight into relationships, we can pose the following hypotheses concerning the interdependence of cognitive load and trust:

1. For a fixed level of trustworthiness, increasing the task complexity (implicitly cognitive load) will affect both the likelihood of a person to rely more heavily on others and the degree of trust they invest in them.

2. For a fixed level of task complexity, varying the trustworthiness of others will affect both the likelihood of the person to rely more heavily on them and hence the degree of cognitive load they perceive during the task.

3. High cognitive load situations are more likely to affect trust judgements that rely on accurate assessments of ability and possibly integrity aspects – since these have been classified as cognitive rather than affective processes during trust judgements.

4. Cultural factors can affect the interdependence of cognitive load and trust, such that cultural biases in trust will be exacerbated under high cognitive load.

Modalities and Data Streams

A number of modalities and data streams were collected in the Australian set of experiment. The experiment was conducted employing dual-task paradigm for higher cognitive load tasks. Subjective ratings of complexity and difficulty were employed after each task set, to ensure that the desired levels of load built into the task design were actually being perceived by the study participants. The Experimental Platform used in the study was developed in-house, that incorporates all data collection, in both versions (high CL and low CL). For details of the study / experiment design and experimental platform, see first year repot for 2011. Following modalities of data were collected:

1. Survey Responses
   - **Pre-Screening Survey**
     A pre-screening survey consisting of 13 questions, with a total of 91 multiple choice questions about the participant's attitudes towards their supervisors and peers, honesty, kindness and trustworthiness, as well as some self-identifying ethnicity and personality based questions.
   - **Mood Survey**
     This single question survey required participants to rate a series of affective aspects, such as happiness and sadness, according to how intensely the feeling was being experienced at the time.
   - **Subjective ratings of mental effort/ task difficulty**
     This single question survey asked participants to rate how difficult the tasks were. It was administered at the end of both the high load and low load sessions. These were collected to ensure that the desired levels of cognitive load were induced.

2. Behavioral Measures
   - **Speech: think-aloud protocols**
     Participants were asked to verbalize their thought processes as they work through the three subtasks. These utterances were recorded.
   - **Justification Text**
     Typing behavior of justification for filling positions, the text provided will be analyzed for temporal and linguistic elements.
   - **Mouse trajectories**
     These are in the form of (x,y) coordinates, and are sampled with enough resolution to reproduce the entire experiment session. The trajectory data will be used to track widget manipulation and log use of the mouse as a placeholder or pointer by hovering over specific areas of the application window. They can also provide an indication of attentional focus.
   - **Other interactive behaviors**
     Application level behaviors such as false starts in answer selections, changes in selections, etc have been collected and will be analyzed.

3. Performance Measures
   - **Ratings, Filling positions and Rankings**:
     The final responses to the actual subtasks.
   - **Time-to-completion**
     Overall and per task.
   - **Performance on secondary task:** Number of notifications correctly added, time-to respond, erroneously added notifications items, errors avoided before adding erroneous items.

5. Analyses

Analysis Plan

The hypotheses described in the earlier section of the same title have been operationalized as follows:

H1:	Participants from a collectivistic culture (e.g. Malaysia) will rate trust higher when applicants have higher benevolence
H2:	Participants from an individualistic culture (e.g.US, Australia) will rate trust higher when applicants have higher ability
H3:	Participants will bin applicants with higher ability in the Supervisor category
H4:	Participants will bin applicants with higher benevolence in the Co-Worker category
H5:	Participants will bin applicants with higher integrity in the Others' Supervisor category
H6:	The above posited cultural effects will be greater under high cognitive load.
H7:	Interactive behaviors, such as speech fluency and mouse trajectories are likely to change during the high cognitive load task when compared to the low load task.

Several analyses are planned to be conducted to test these hypotheses. First, the survey data will be aggregated based on the pre-established scales used. Reliability analyses will be conducted to ensure that these measures are reliable. Various analysis techniques (e.g., *t*-tests,

ANOVA, regression tests) will be used depending on the hypothesis to be tested. Principle component analysis will be needed for the survey and questionnaire answers.

Following table summarizes various categories of analyses planned for the data:

Analysis category	Types in each category	Description
Subjective Ratings Analysis		To validate the experiment design for required task difficulty / mental effort.
Linguistic Analysis of Speech Data	Pause Analysis Linguistic Category Analysis Language complexity analysis	To analyze the speech and linguistic behavioral changes for various cognitive load and trust conditions.
Signal Analysis of Speech Data	Analysis of pitch, tone, speech rate, intensity, energy and other speech signal features.	To analyze the variations in speech signal patterns for different cognitive load and trust conditions.
Interaction Data Analysis	Mouse interaction analysis Keyboard interaction analysis	To analyze trajectories and typing behavior and their temporal and linguistic elements.
Performance Analysis	Performance on Ratings, Filling positions and Rankings. Time-to-completion - Overall and per task. Performance on dual-task: Number of notifications correctly added, missed candidates, time-to respond, erroneously added candidates	The final responses to the actual subtasks. To analyze the performance variations under various load and trust conditions.
Qualitative Analysis of Speech Data	Analysis of the thought-process through the speech and transcriptions	To understand the thought-process through which people make trust judgments under different cognitive load situations.

A number of features of interest are being extracted and annotated. The following details some of the feature extraction activities being carried out on each of the behavioral measures recorded.

Speech data

1. Data cleaning (e.g. remove cross-talk, segmentation)

The speech data has been recorded in segments, which correspond to each of the three subtasks. Since the experiments took place in a classroom laboratory, a number of participants completed the sessions at the same time. Although directed microphone headsets were use, and participants were seated as far away as possible from one another, there is a chance that cross-talk has affected the speech recordings. It will be necessary to clean the data by extracting any noise or speech from other participants from the recording.

2. Build CL models, test data

Some of the data will be used to create low load and high load models of speech for each of the three tasks, while the rest of the data will be used to test the models. This will verify whether cognitive load can be detected from the acoustic features of speech in this application.

3. Linguistic analysis of think-aloud speech

Once the speech data is pre-processed and cleaned, mid-level features such as pause frequencies and lengths can also be annotated. Linguistic speech features can also be collected from the transcripts (which themselves could be generated automatically). Other features such as frequency and type of pronoun use, sentence complexity (including sentence length and average word length), total text length, use of affective words, use of cognitive words, among other categories, will also be examined.

4. Transcriptions and qualitative analysis of speech data

Finally, qualitative analysis can be useful in this instance to further understand the thought process through which the participant arrives at their response. Similarities in thought processes between questions/sub-questions can provide more information about how trust judgments are made.

Justification Text

The justification text will undergo linguistic analysis, including: frequency and type of pronoun use, sentence complexity (including sentence length and average word length), total text length, use of affective words, use of cognitive words, among other word categories. These features will be used for comparison purposes between the low and high load conditions.

Mouse trajectories

Initially, a parsing tool will be built that can display each trajectory along a time scale, and allow closer inspection of movement. This will allow exploratory analysis/ inspection of mouse behaviors which are typical of this application. Some basic features that can be automatically extracted from this dataset include:
- Time spent moving mouse
- Distance traveled per task/ per session
- Categorizing time spent in different screen/window areas on a per-task basis
- Which areas of the screen were most frequented
- How much time spent on specific widgets, e.g. drop down boxes.
- Which information was looked at when answering which questions.
- Which questions were hesitated on/ Which questions they were much more decisive on
- Pauses in mouse movement indicate thinking – this will help to identify sections of high load.

While there may be large individual differences, the trends may still indicate relative changes at different points in time during the task.

## Other interactive behaviors

Application level behaviors, such as false starts in answer selections, changes in selections, etc. can also give an indication of high load instances within the session or task.

Analyses Results

As mentioned earlier, this year, milestones 5-10 have been amended from previous documents to include some analyses of more interactive features as well as due to a team member from NICTA who left the project, the project's timeline was affected. We have already conducted some partial analyses so far of the data collected from the Australian site and the results of those analyses are being discussed in the following. As of the time of this writing, the project is running on time and we will be completing all the analyses planned above as per the schedule and detailed in future report.

Subjective ratings of mental effort/ task difficulty

To validate the experiment design for required cognitive load levels the subjective ratings of mental effort or task difficulty were collected from the participants. These were collected at the end of both the high load and low load task sessions and were based on a seven-point Likert scale (from 1="Extremely easy" to 7="Extremely difficult"). The analysis of the subjective ratings showed a mean ratings of 3.625 for high cognitive load condition and 3.037 for low load condition as shown if the following graph and statistically significant difference between them ($t(72)=5.201$, $p<.001$). This confirmed the effective experiment design inducing the required levels of task difficulty and/or cognitive load levels as expected.
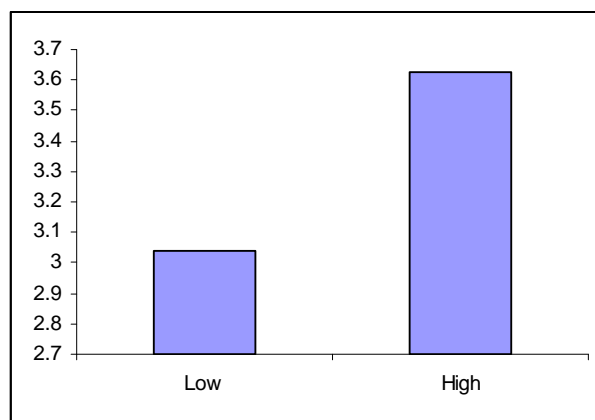
Participants' Subjective Ratings of Task Difficulty

Linguistic Analysis of Speech Data

We have already completed partial linguistic analysis of the think-aloud speech recorded from the participants. The linguistic analysis is being carried out in three different areas; pause analysis, linguistic category analysis, and language complexity analysis; the objective was to analyze the linguistic behavioral changes for various cognitive load and trust conditions.
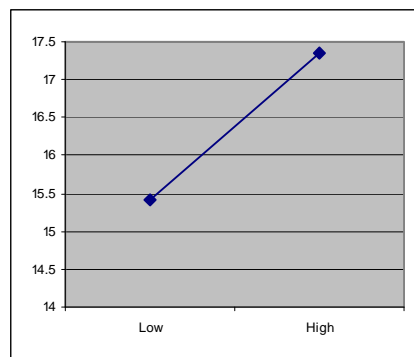
1. Pause Analysis

The mid-level speech features such as pause frequencies and their lengths etc were analyzed. Fifteen different pause features were analyzed, which are listed in the following table:
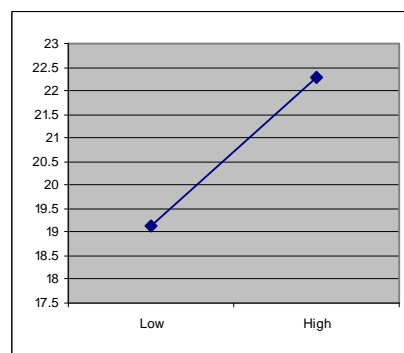
Pause Features	Description
Average response latency	In seconds
# silent pauses	Frequency of silent (voiceless) segments
# filled pauses	Frequency of filled (voiced) segments, e.g. ahhh, umm.
# total pauses	Freq. of total pauses
ag # silent pauses/min	Average frequency of silent pauses per minute (normalized)
ag # filled pauses/min	Average frequency of filled pauses per minute (normalized)
avg # pauses/min	Average frequency of total pauses per minute
avg silent pause length	Average length of silent pauses (in seconds)
avg filled pause length	Average length of filled pauses (in seconds)
avg total pause length	Average length of total pauses (in seconds)
% of total time pausing	Percentage of total time in pausing
avg # hesitations	Average frequency of hesitations
avg # self-corrections	Average frequency of self-corrections
avg # incomplete sentences	Average frequency of using incomplete sentences
avg # repetitions	Average frequency of repetitions

Various hypotheses related to these pause features were formed with regard to their behavior under low vs. high cognitive load conditions and statistical tests (including paired sample *t*-tests) were performed. Generally, as per many previous studies [2, 3], it was expected that participants will use more and longer pauses under high cognitive load condition as compared to low cognitive load one.

The pause features were manually annotated using the ELAN annotation tool [4]. Because there are over 70 participants, whose speech is being annotated, we have been able to complete only 10 participants' speech annotations. Therefore, the tests so far have failed to show any significant results, but the trends so far are in the expected directions. Following graphs show some pause feature trends we have got so far. We expect that these trends (and trends for other features) will persist and show statistically significant differences with enough power, once all the participants' speech has been annotated.
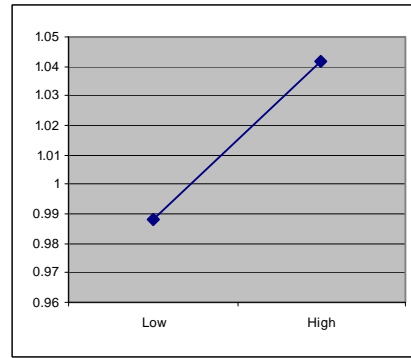

Avg Freq of Silent Pauses per Minute   Avg Freq of Total Pauses per Minute
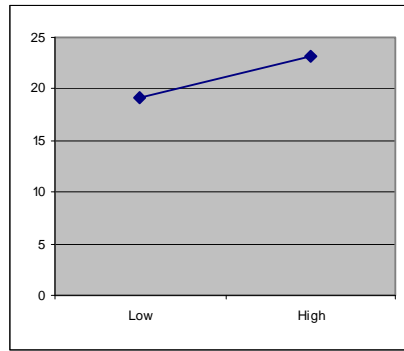
Percentage of time pausing (%)


Avg Length of Pauses (seconds)

2. Language Category Analysis

The language category analysis involved examination of different types or categories of words used by the participants in their think-aloud speech under the two cognitive load conditions and various trust situations. Following table lists those word categories that were selected for our analysis based on their relevance from the literature [5-7].

Linguistic Category	Example words
Word Count	
Words per minute of speech	
Words per sentence	
Long words (words >6 letters)	
Avg. # of sentences	
Personal pronouns	I, they, her, we
Impersonal pronouns	it, those, it's, that
Adverbs	very, really, quickly, mostly
Negations	no, not, never, neither
Quantifiers	few, many, much, fairly
Swear words	damn, shit, fuck, piss
Affective (emotional) processes	happy, cry, glad, afraid
Positive emotions	nice, sweet, cool
Negative emotions	ugly, nasty, bad, fail, sorry
Anxiety	worried, fearful, nervous
Anger	hate, kill, annoyed
Sadness	sad, grief, cry
Cognitive processes	know, cause, opinion
Insight	think, know, consider
Causation	hence, effect, because
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never, absolutely
Achievement	win, hero, ability, perform
Assent	agree, ok, yes, cool
Trust	trust, believe, sure
Distrust	doubt, disbelieve, suspicious

In order to extract these linguistic features, we needed speech transcriptions of the participants' spoken speech. We initially planned to use some automatic speech recognition (ASR) system to generate the transcriptions automatically. After our effort with one of a "robust" speech-to-text (STT) systems, we realized that without proper training of the ASR system, which is hugely time-consuming task (and is not possible to do with our type of experiment/study), the STT performance will not generate transcriptions good enough for our purpose. Hence, for the sake of our analyses, we had to manually transcribe and annotate the participants' speech using the ELAN tool. So far the transcriptions are still under process and we have only been able to analyze 10 participants' transcription data. The language category features listed above were automatically extracted from the transcripts using a linguistic analysis tool called LIWC2007 [8], which extracted most of these features as percentage of total words spoken by a participant.

Various hypotheses related to these language category features were formulated with regard to their behavior under low vs. high cognitive load conditions and statistical tests (including paired sample t-tests) were performed. Once again due to a fewer number of transcriptions available so far, the tests have failed to show any significant differences, but many of these features show the trends as expected. Following graphs show some linguistic feature trends. We expect that these trends will persist and show statistically significant differences, once all the participants' speech has been transcribed.
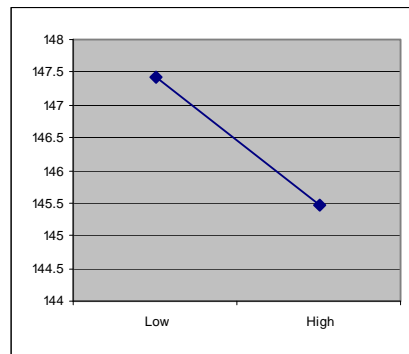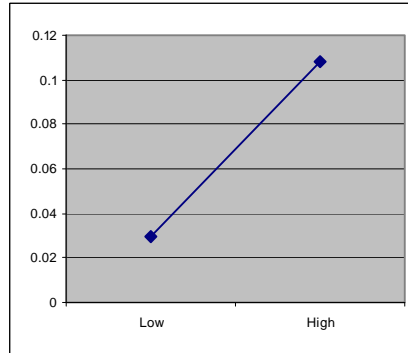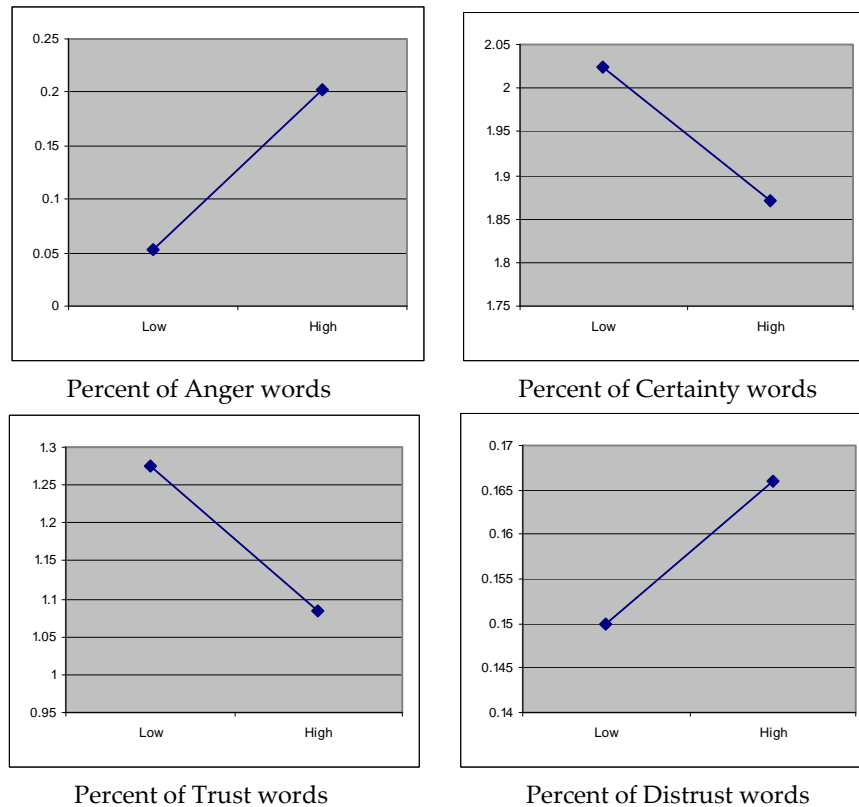

Average words per minute
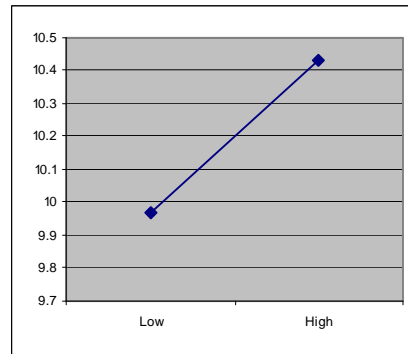
Average words per sentence
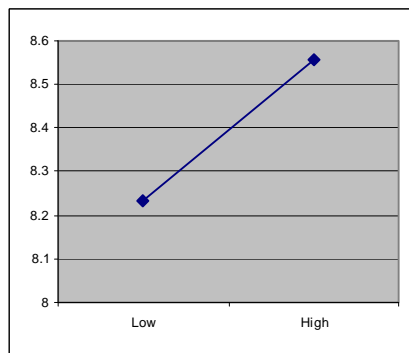
Percent of Negative emotion words

Percent of Swear words

Percent of Anger words

Percent of Certainty words

Percent of Trust words

Percent of Distrust words

6. Data Collection

Schedule

The study and data collection schedule has slightly changed due to some unforeseen reasons including arrangement of participants. Nevertheless, the Australian data was collected as per original schedule in 2011 and about 100 students from the University of Sydney participated in the user study on the 18th October.

As of this writing, data collection at Malaysian site is currently under progress. The US site is still planning for the data collection and roughly 160 students are expected to participate in June 2012.

Data Collection Summary: Australia

The Australian group's data has already undergone preliminary analysis to determine the quality of the data collected and validation of the protocol. It is expected that no change will be required in the software tool for other sites. If the new version of the tool is substantially different from the version administered at Australia due to confounds or other issues, a new set of students can be canvassed from University of Sydney this year to complete the new version of the user study. Some statistics of the Australian data collection are as following:

- 91 subjects completed both conditions (high and low CL)
- Approximately 239 survey/response data points per subject
- Speech data: 6.5Gb = 58 hours of speech
- Interactive Behaviour: ~96 million data points including mouse trajectories, selection, typing, browsing activity (attentional focus)

Data Collection Summary: Malaysia

As of this writing, about 80 university students are currently participating in the user study. The study administration and data collection process is running smoothly. More details will be provided in future report as the process completes and information becomes available. Other details on the activities conducted by the Malaysian site can be found in the project companion annual report [9].

Data Collection Summary: US

As of this writing, US data collection is still being planned and is schedule to run in June 2012. More details will be provided in future report as they become available.

7. Operational Processes

IRB Approvals

Dr. Asif Khawaja has been added to the IRB documentation as part of joining the team on the Australian side. All Australian team members (Fang Chen and Asif Khawaja) have completed refresher CITI training and have received their certificates.

8. Conclusion

In conclusion, we have summarized the second year research activities as part of the "Effects of Cognitive Load on Trust" project in conjunction with the US AFRL and Sunway University, Malaysia. NICTA's role comprised the measurement and assessment of cognitive load through speech, linguistic, and other interaction modalities. The second year of the project was dedicated mainly to the analysis of the Australian dataset, collected in 2011, and preparing for the second data collection phase from Malaysian and the US sites. As of this writing, the Malaysian data collection is underway and the US site is planning its data collection activity and will be running the experiment in June 2012.

An updated project plan was presented along with a description of various modalities and data streams to be analyzed for this research including subjective ratings, speech signal data, linguistic data, and interaction data. A multidimensional data analysis was planned to analyze the multimodal data collected from Australian site and their behavior under different cognitive load conditions. We have already conducted detailed analysis of some of the data collected including the subjective ratings of mental effort and linguistic analyses of the speech data that included pause analysis of various pausing features and language category analysis of various linguistic category features.

The primary outcomes of these analyses were also presented showing that participants rated the high cognitive load tasks as requiring more mental effort. The speech results showed when interacting with a complex system and performing a high cognitive load task, people tend to pause more and longer as compared to low cognitive load tasks. The preliminary results also showed that people tend to use various types of words differently under different cognitive load situations. Specifically, it was observed that under high load conditions, people used longer sentences, more negative emotion words, more swear words, more anger words, and fewer trust words and more distrust words.

More data analysis is planned for these and several other modality data streams in the third year of the research. Detailed findings on how cognitive load affects people's behavior and their trust perception will be presented in the Final year annual report.

9. References

[1]   R. C. Mayer, et al., "An Integrative Model of Organizational Trust," The Academy of
        Management Review, vol. 20, pp. 709-734, 1995.
[2]   M. A. Khawaja, et al., "Think before you talk: An empirical study of Relationship
        between Speech pauses and Cognitive load," presented at the Australasian
        Computer-Human Interaction Conference (OzCHI'08), Cairns, Australia, 2008.
[3]   A. Berthold and A. Jameson, "Interpreting Symptoms of Cognitive Load in Speech
        Input.," presented at the Seventh International Conference on User Modeling (UM99),
        1999.
[4]   "ELAN Linguistic Annotator," ed: Max Planck Institute for Psycholinguistics, The
        Netherlands, Last accessed: May 2012.
[5]   M. A. Khawaja, et al., "Using Language Complexity to Measure Cognitive Load for
        Adaptive Interaction Design," presented at the Proceedings of International
        Conference on Intelligent User Interfaces (IUI 2010), Hong Kong, China, 2010.
[6]   M. A. Khawaja, et al., "Cognitive Load Measurement from User's Linguistic Speech
        Features for Adaptive Interaction Design," presented at the Proceedings of
        International Conference on Human-Computer Interaction (INTERACT'09), Part I,
        LNCS 5726, Uppsala, Sweden, 2009.
[7]   M. A. Khawaja, et al., "Analysis of Collaborative Communication for Linguistic Cues
        of Cognitive Load," International Journal of Human Factors and Ergonomic Society,
        vol. http://dx.doi.org/10.1177/0018720811431258, January 2012.
[8]   J. W. Pennebaker, et al., "The Development and Psychometric Properties of
        LIWC2007," 2007.
[9]   M. Lin and A. Yeo, "Influence of Cultural Cognition, Social Aspect of Culture, and
        Personality on Trust", First Year Annual Report (AOARD-104129), 2011.